GAN-Based Image Enrichment in Digital Pathology Boosts Segmentation Accuracy

Laxmi Gupta[1], Barbara M. Klinkhammer[2],
Peter Boor[2], Dorit Merhof[1], and Michael Gadermayr[1,3]

[1] Institute of Imaging & Computer Vision, RWTH Aachen University, Aachen, Germany
[2] Institute Pathology, University Hospital Aachen, RWTH Aachen University, Aachen, Germany
[3] Salzburg University of Applied Sciences, Salzburg, Austria

Abstract. We introduce the idea of 'image enrichment' whereby the information content of images is increased in order to enhance segmentation accuracy. Unlike in data augmentation, the focus is not on increasing the number of training samples (by adding new virtual samples), but on increasing the information for each sample. For this purpose, we use a GAN-based image-to-image translation approach to generate corresponding virtual samples from a given (original) image. The virtual samples are then merged with the original sample to create a multi-channel image, which serves as the enriched image. We train and test a segmentation network on enriched images showing kidney pathology and obtain segmentation scores exhibiting an improvement compared to conventional processing of the original images only. We perform an extensive evaluation and discuss the reasons for the improvement.

Keywords: histology, kidney, augmentation, enrichment, sensor fusion, segmentation, adversarial networks

1 Motivation

Data augmentation has become common knowledge and exhibits an indispensable method to boost the performance of state-of-the-art machine learning approaches, especially if the amount of training samples is relatively small. While data augmentation typically refers to increasing the number of samples, we introduce image enrichment for increasing the information content for every individual sample. An application, for example, is given by medical image fusion [5], where the information of one image (e.g. CT) is enriched by information obtained from another image showing the same underlying structure, but using a different imaging modality (e.g. MRI). The additional information obtained by merging the images either facilitates medical diagnosis or exhibits the basis for further computer-aided decision support systems.

While in specific fields multi-modal image data is available and indispensable for reliable diagnosis, in many other areas we (need to) deal with single images and mostly do not even think about adding data from a different domain. However, recent achievements in deep-learning facilitate translations between different imaging modalities, e.g. between CT and MRI scans [8]. A conversion is obtained by the so-called image-to-image translation approaches, showing attractive and realistic output [4, 9]. A specific

generative adversarial network (GAN) architecture also enables translation without the need for any corresponding pairs to train the networks [9]. The proposed GANs rely on cyclic loss functions which are combined with the GAN loss. The cyclic loss ensures that after circular translations (e.g. from a domain A to a domain B and back to domain A) the final images are similar to the input image. The GAN loss ensures that the output images look real. As image pairs often cannot be obtained (or are at least difficult and/or expensive to achieve), this architecture, allowing unpaired training, opens up entirely new opportunities for the field of medical image analysis.

In the field of digital pathology, GANs can also be used for image normalization [1, 3] and for domain adaptation [2]. The authors of [2] considered a scenario, where labeled samples are only available in a source stain and not in the target stain, and performed domain adaptation on image-level using image-to-image translation. They concluded that GANs can be applied to create realistic fake images showing a stain different from the input stain. They showed that the fake images can be effectively used for segmentation in a domain adaptation scenario but noticed that the direction of translation makes a clear difference. A translation to one specific stain showed higher segmentation scores than to others. They assume that certain stains are easy to segment while others more difficult. To obtain optimum results, they suggest to translate to the stain which is easier to segment (either source or target stain) before segmentation.

Contribution

With the outlook of improving segmentation accuracy, we propose a method called 'image enrichment', whereby the information content within each sample is increased. We consider a scenario where labeled images (training and testing images) are available for a single stain only. Unlike in recent work [2], we do not consider a domain change scenario, but focus on the segmentation of samples showing the same stain as the training data. We perform image-to-image translation to generate virtual images showing stains different from the input sample. The corresponding virtual images are merged with the original samples and further utilized for training and testing the segmentation network. In an experimental study on kidney pathology data, we show that this approach is capable of increasing the segmentation accuracy and discuss the underlying causes. Related work [7] presents a similar generator-to-classifier network that jointly optimizes stain-translation and classification. In contrast, we investigate an approach based on two individual networks and focus on a segmentation task. Our two step approach allows higher flexibility at the time of application.

2 Methods

We consider a domain of labeled images, denoted as S_0, referring to one specific stain, and further domains S_1, ..., S_n of images, each showing a specific stain different from S_0. For the domains S_1, ..., S_n, no labels are available. There is also no need for any corresponding image pairs. The only restriction is that all sets show similar underlying tissue (e.g. kidney tissue).

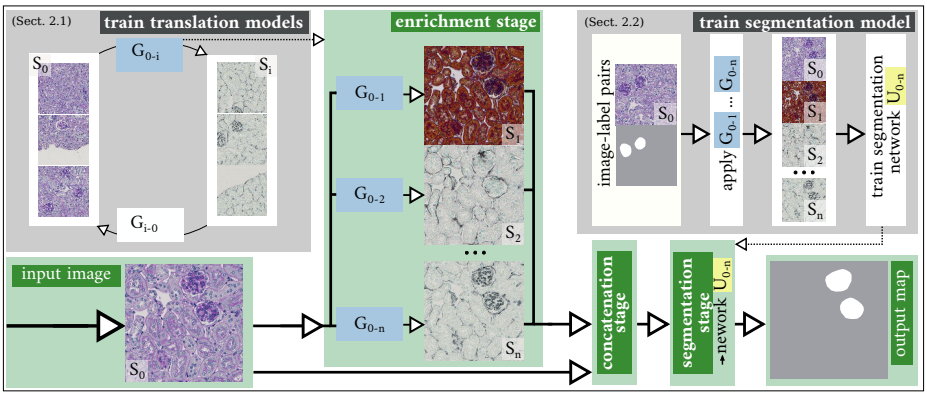


Fig. 1: Outline of the proposed segmentation pipeline: for an image to be segmented (input image), first corresponding virtual images are created (enrichment stage) and concatenated (concatenation stage). Finally, the resulting multi-channel image is segmented using a segmentation network (segmentation stage). The image translation and segmentation models are trained individually.

We propose a segmentation pipeline consisting of an enrichment, a concatenation, and a segmentation stage, as shown in Fig. 1. An input image of domain S_0 to be segmented is first passed to the enrichment stage. In this phase, several generators G_{0-1}, G_{0-2}, ..., G_{0-n} are applied to the input image in order to generate a virtual sample for each of the domains S_1 to S_n. The networks G_{0-i} are trained beforehand in an unpaired manner using a cycle-consistency GAN. Next, the input image is concatenated with all the virtual images generated from it. The resulting domain is referred to as $S_{0,1,\ldots,n}$, in the following also as S_{All}. Supposing that the original and the n virtual images consist of $M \times N$ pixels and C color channels, we obtain multi-channel images exhibiting a dimensionality of $M \times N \times (C \cdot (n+1))$. Finally, the multi-channel images so obtained are fed to a fully-convolutional network for segmentation. The rationale behind this approach is that image translation in histology effectively generates highly realistic images [2]. Here, we assess if the additional information, available from a set of different histological stains, is helpful to support the final network to learn the segmentation task. Each trained generator ($G_{0-i} : S_0 \rightarrow S_i$) is only based on the input image and does not incorporate any further information when creating virtual samples. However, additional information is created due to the learned ability of the generator to convert between the different domains. It can be argued that segmentation should be independent of the modality because the underlying image content in all the domains is the same. Nevertheless, it was shown that the domain in which segmentation is performed does influence the accuracy [2].

2.1 Image Translation Model

Each of the translation models G_{0-1}, G_{0-2}, ..., G_{0-n} are trained individually. To train an individual image translation model G_{0-i}, firstly, patches from the input image

domain S_0 and the target domain S_i are extracted from the original whole slide images (WSIs). Patch extraction is needed because due to the large size of the WSIs (few gigapixels), a holistic processing of complete images is not feasible. For training, patches with a size of 512 $\times$ 512 pixels are extracted from the original WSIs. The patches utilized for training are uniformly sampled. A non-uniform sampling in both domains focusing more on positive objects as suggested in [2] is not possible in this unsupervised image translation scenario (as labels are only available in one domain). With these patches, a cycleGAN [9] consisting of two generative models, $F : \mathcal{X} \rightarrow \mathcal{Y}$ and $G : \mathcal{Y} \rightarrow \mathcal{X}$ and two discriminators D_X and D_Y is trained. The losses include the established GAN-loss $\mathcal{L}_{GAN}$, the cycle-loss $\mathcal{L}_{cyc}$, and the identity loss $\mathcal{L}_{id}$ (with corresponding weights $w_{GAN} = 1, w_{cyc} = 1, w_{id} = 1$). Particularly, the cycle-loss

$$\mathcal{L}_{cyc} = \mathbb{E}_{x \sim p_{data}(x)}[||G(F(x)) - x||_1] + \mathbb{E}_{y \sim p_{data}(y)}[||F(G(y)) - y||_1] \,, \quad (1)$$

forces the generator to maintain the image content without requiring paired samples. The identity-loss is applied to stabilize the training process as suggested in [9]. For training, standard data augmentation (rotation, flipping) is applied. Apart from a U-Net generator network [6], the standard configuration based on the patch-wise CNN discriminator is used [9][1]. Initial learning rate is set to 0.0002. Adam optimizer is used.

2.2 Segmentation Model

For segmentation, we rely on an established fully-convolutional network architecture, specifically the so-called U-Net [6] which was successfully applied for segmenting kidney pathology [2]. For taking the distribution of objects into account (the glomeruli are small, sparse objects covering only approximately 2% of the renal tissue area) training patches (512×512 pixels) are not randomly extracted. Instead, as suggested in [2], 50% of the patches are extracted from regions containing objects (to obtain class balance) whereas the other 50% are randomly extracted (to include regions far away from the objects-of-interest). Batch-size is set to one and L2-normalization is applied. Due to the very low within-stain variability in the data set, there is no need to apply stain-normalization. Standard data augmentation (rotation, flipping) is applied. Initial learning rate is set to 0.001, and Adam optimizer is used.

3 Image Data & Experimental Settings

We investigate WSIs showing tissue of mouse kidney. The images are captured with the whole slide scanner model C9600-12, by Hamamatsu with a 40$\times$ objective lens. The overall data set comprises: 23$\times$ periodic acid Schiff (PAS), 12$\times$ Acid Fuchsin Orange G (AFOG), 12$\times$ cluster-of-differentiation (CD31) stained WSIs, and 12$\times$ images dyed with a stain focused on highlighting Collagen III (Col3).

As suggested in [2], we perform both, segmentation and image translation, on the second highest resolution (20$\times$ magnification). We consider a scenario where manually annotated PAS stained WSIs are available for training a supervised segmentation model.

[1] We use the PyTorch reference implementation [9].

Therefore, PAS is considered as the S_0 domain. Consequently, we refer to CD31 as S_1, AFOG as S_2, and Col3 as S_3. Eleven of the PAS images and all of the CD31, AFOG, and Col3 images are used for training the translation model, where the three latter stains are employed for image enrichment. We train the segmentation model on patches extracted from the WSIs (400 from each WSI). However, to avoid any bias, the training and testing data is separated on WSI level. For this purpose, we randomly select ten WSIs for training and two for testing. This procedure is repeated 12 times.

4 Evaluation Metrics

We compute precision, recall and the Dice similarity coefficient (DSC) individually for each of 12 repetitions. We investigate the baseline setting with the original PAS images (S_0) and the proposed method using additionally all the virtual stains (S_{All}). To gain further insight, we also evaluate combinations of the PAS domain with single virtual domains ($S_{0,1}$, $S_{0,2}$, $S_{0,3}$) and also individual virtual domains without any real data (S_1, S_2, S_3). Two setups are investigated.

Setup 1

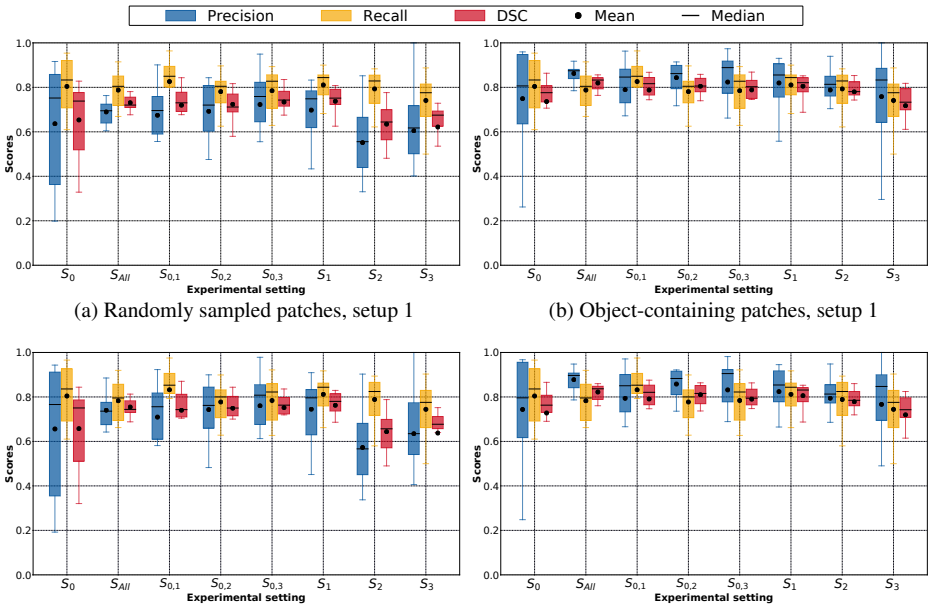
Evaluation is performed for (a) randomly sampled patches representing the DSC on WSI level, and (b) patches containing (at least one pixel of) one or more objects. The latter is motivated by the fact that large regions do not contain any objects and can be easily manually "excluded".

Setup 2

In this setting, we incorporate the fact that small objects (specifically objects with an area below 5,000 pixels) are not relevant for further analysis. These small objects occur in 2D images if a 3D glomerulus is cut marginally (i.e. the cut is close to the object's border). As it is difficult to determine the exact size during manual annotation, these objects are partly labeled in the ground truth. Consequently, we additionally compute the DSC as follows. If a small object (area < 5000 pixels) is detected but not labeled in the ground truth, it is ignored when calculating the measures (precision, recall, DSC). If a small object is marked in the ground-truth and not detected, it is also ignored.

5 Results

Fig. 2 shows the experimental results. The subfigures correspond to the four evaluation methods (a)–(d) explained above. Results are shown for the baseline performing segmentation on original PAS images (S_0), the proposed method incorporating all the available stains (S_{All}), combinations of real PAS and one virtual stain, and individual virtual stains. In all the four subfigures, the proposed method S_{All} shows higher mean DSCs ((a): 0.73 vs. 0.65, (b): 0.82 vs. 0.74, (c): 0.75 vs. 0.66, (d): 0.82 vs 0.73) and lower standard deviations, as compared to the baseline approach S_0. Especially precision is increased on average while the standard deviation in clearly reduced. A similar



(a) Randomly sampled patches, setup 1

(b) Object-containing patches, setup 1

(c) Randomly sampled patches, setup 2

(d) Object-containing patches, setup 2

Fig. 2: Segmentation scores (precision, recall, DSCs) shown individually for the baseline method (S_0), the proposed method (S_{All}), the concatenation of original images with a single virtual image ($S_{0,1}$, $S_{0,2}$, $S_{0,3}$), and the individual virtual images (S_1, S_2, S_3). The subfigures correspond to the evaluation settings Setup 1 and Setup 2.

trend is observed for the other test settings, namely $S_{0,1}$, $S_{0,2}$, and $S_{0,3}$, also showing improvements compared to S_0 and reaching scores similar to S_{All}. Regarding the settings based on a single virtual stain only, we observe divergent scores. S_2 and S_3 show partly degraded performance in case of randomly samples patches. Setting S_1 shows segmentation scores similar to S_{All} and clearly higher than the baseline S_0. It is noteworthy that precision (and hence DSC) increases when we consider only object-containing patches (Fig. 2(b)). The difference is especially relevant for the settings S_1, S_2, and S_3. Setup 1 and setup 2 show similar trends. In Fig. 3, we see examples of the segmentation outputs for the baseline and the proposed setting in comparison with the ground truth. Fig. 4 shows qualitative results of the image translation process. We do not notice any systematic differences between virtual and real images. The virtual patches also show a high correspondence (i.e. the object outlines do not change) to the real images.

6 Discussion

With the help of unpaired image-to-image translation, we propose a method which solves a segmentation task in two steps. In the first step the image data is enriched and in the second step, the available data is fed into a specifically trained segmentation

network. Summarizing the results, we notice two core findings. Firstly, the proposed method including all available virtual stains (S_{All}) exhibits improved scores for all evaluation settings compared to the straight-forward segmentation of S_0 (PAS) images. Also experiments with subsets using only S_0 and one single virtual stain consistently show improved (mean) DSCs. Secondly, we notice that especially one virtual stain, namely S_1, exhibits scores similar to the setting S_{All} with all information merged. Highly interestingly, the virtual domain S_1 is obviously even "easier" to segment than the original domain S_0. Based on these two findings, we summarize that image enrichment based on image-to-image translation can actually improve the segmentability of images. In other words, solving segmentation in two steps can be more accurate than performing a segmentation in a single step. We explain this as follows: for the first step, we train a network to perform a task on a rather low level. To perform image translation in histology, details are more important than large contextual information. In the second step, large context is required to decide if a potential object is a real object or a similar artifact. We hypothesize that two networks, individually specialized and trained for different sub-problems, are more powerful than one single network applied to one, but more difficult task. However, this requires that an appropriate intermediate domain (in our case, an appropriate histological stain) exists.

Regarding the segmentation of single (virtual or real) and combined domains, we summarize that a single highly powerful domain (S_1) is sufficient in our setting to show superior results. A fusion with further virtual or real stains does not show any further improvements. Interestingly, we find that the virtual domains show decreased standard deviations which is highly likely due to the lower variability within the virtual stains compared to the real PAS stain. This might be due to a naturally low variability of the histological stains (considered as virtual stains) or due to a normalizing effect of the generator network. Even though the slides have a homogeneous apprearance in terms of color, we do observe intra-slide variations in texture. Another positive effect could also be introduced by the generator network which might be able to compensate for degradations in the image domain. For example, the network might remove minor artifacts and thereby generate optimized virtual domains. However, we did not notice such cases and it can also be argued that the generator is not optimized to remove artifacts, but rather to produce samples similar to those drawn from the original distribution. In order to clearly separate the effect of the individual properties of the

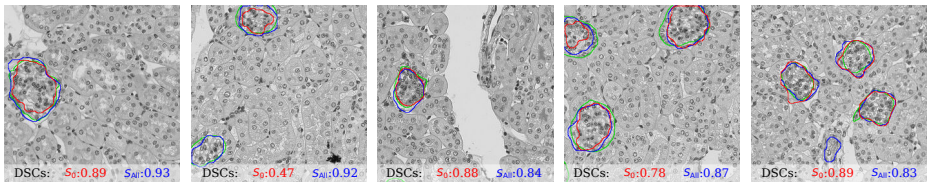


Fig. 3: Qualitative Results of the segmentation process: Red and blue show the segmentation outputs of the approaches based on S_0 and S_{All}, respectively. Green shows the ground truth. To improve visibility, we show two approaches only and provide an overlay with gray-scale images.

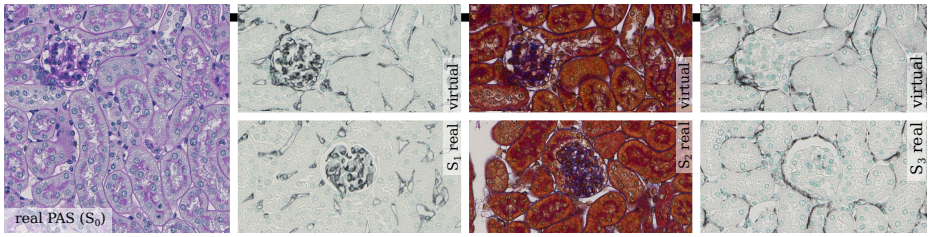


Fig. 4: Example images showing a real S_0 (PAS) patch, corresponding virtual images (top row) and similar real images (bottom row) for comparison. While the virtual images correspond to the real patch, the real ones surely do not.

underlying histological stain and the potentially normalizing effect of the generator network, further experiments need to be performed in the future. This requires large amounts of annotated samples for each of the stains used as virtual stains as well, for supervisedly training the segmentation network.

To conclude, we propose the idea of image enrichment, which exploits the fact that image-to-image translation based on unpaired training can be used to improve segmentation accuracy. In a two-step approach, we firstly enrich, i.e. boost the information content on the available data using unpaired image-to-image translation, and secondly, train a segmentation network which benefits from this enriched data. We tested segmentation performance on a given stain by enriching it with three additional stains and obtained improved segmentation scores. Even with a specific single virtual stain, we achieved improved scores, similar to the DSCs obtained when merging all available data. This is expected to be due to the fact that specific histological stains are more appropriate for subsequent segmentation. Secondly, the generator might be able to improve average image quality by reducing variability (and artifacts) in the image domain. Noticing generally lower standard deviations for the virtual stains and stain specific differences, we strongly assume that both effects play a vital role. To provide a clear answer, further experiments based on additional labeled training data will be performed in the future.

Acknowledgment: This work was supported by the German Research Foundation (DFG) under grant no. ME3737/3-1.

References

1. Bentaieb, A., Hamarneh, G.: Adversarial stain transfer for histopathology image analysis. IEEE Transactions on Medical Imaging **37**(3), 792–802 (2018)
2. Gadermayr, M., Appel, V., Klinkhammer, B.M., Boor, P., Merhof, D.: Which way round? a study on the performance of stain-translation for segmenting arbitrarily dyed histological images. In: Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'18). pp. 165–173 (2018)
3. Ghazvinian Zanjani, F., Zinger, S., de With, P.H.N., E. Bejnordi, B., van der Laak, J.A.W.M.: Histopathology stain-color normalization using deep generative models. In: Proceedings of the Conference on Medical Imaging with Deep Learning (MIDL 2018). pp. 1–11 (2018)

4. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR'17) (2017)
5. James, A.P., Dasarathy, B.V.: Medical image fusion: A survey of the state of the art. Information Fusion **19**, 4–19 (2014)
6. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer Aided Interventions (MICCAI'15), pp. 234–241 (2015)
7. Wu, B., Zhang, X., Zhao, S., Xie, L., Zeng, C., Liu, Z., Sun, G.: G2c: A generator-to-classifier framework integrating multi-stained visual cues for pathological glomerulus classification. ArXiv **abs/1807.03136** (2019)
8. Zhao, Y., Liao, S., Guo, Y., Zhao, L., Yan, Z., Hong, S., Hermosillo, G., Liu, T., Zhou, X.S., Zhan, Y.: Towards MR-only radiotherapy treatment planning: Synthetic CT generation using multi-view deep convolutional neural networks. In: Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI'18). pp. 286–294 (2018)
9. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the International Conference on Computer Vision (ICCV'17) (2017)